HOOFDSTUK 9 - VAN REGELGEBASEERDE SYSTEMEN NAAR MACHINE LEARNING

1. LEERPLANDOELEN KathOndVla Leerplan Communicatiewetenschappen LPD 7+ De leerlingen illustreren mogelijkheden en beperkingen van taaltechnologische toepassingen.

1. Beantwoord de volgende meerkeuzevragen.

1) Wat is een chatbot?

Een chatbot is de algemene naam voor software die gesprekken met mensen kan voeren. In de meest eenvoudige vorm typen gebruikers hun zinnen in en verschijnen de reacties als tekst op het scherm. Sommige chatbots maken gebruik van spraaktechnologie, zodat ze ook gesproken tekst kunnen interpreteren en zelf hun antwoord kunnen voorlezen.

2) Welke van de volgende toepassingen maakt waarschijnlijk GEEN gebruik van taaltechnologie?

- Spamfilters in e-mailprogramma’s maken gebruik van taaltechnologie om verdachte e-mails te herkennen, onder andere op basis van verdachte woorden, zoals ‘prijs’, ‘gratis’ of ‘gewonnen’.

- Zoekmachines maken gebruik van taaltechnologie om informatie te vinden die aansluit bij de zoekopdracht van de gebruiker. Zo houden ze bijvoorbeeld rekening met vervoegingen, verbuigingen en synoniemen. Geavanceerde zoekmachines op het internet maken ook gebruik van taaltechnologie, om de kwaliteit van de gevonden websites te beoordelen. - Automatische ondertiteling gebruikt een combinatie van taal- en spraaktechnologie. Het zet gesproken taal niet alleen om in geschreven tekst, maar zorgt er ook voor dat de geschreven zinnen grammaticaal zijn en niet te lang worden.

- Hoewel automatische handschriftherkenning te maken heeft met het herkennen van geschreven tekst, is het voornamelijk een vorm van beeldverwerking in plaats van taaltechnologie. Het richt zich op de identificatie van vormen en patronen in geschreven tekens, om die om te zetten in digitale tekst, zonder noodzakelijkerwijs de betekenis of structuur van de taal te verwerken.

3) Hoe verhoudt taaltechnologie zich tot artificiële intelligentie? - Taaltechnologie maakt gebruik van technieken uit de artificiële intelligentie (AI) om taal te verwerken en te begrijpen.

- Taaltechnologie is een tak van AI die zich bezighoudt met de interactie tussen computers en menselijke taal.

- AI is een breder vakgebied dat taaltechnologie omvat samen met andere subvelden, zoals bijvoorbeeld beeldherkenning en robotica. - Taaltechnologie is specifiek gericht op de verwerking van taal, terwijl AI als geheel een breder

spectrum van toepassingen omvat. **4) Wat maakt taaltechnologie zo uitdagend?**
- De grote variëteit aan talen en dialecten is een belangrijke uitdaging voor taaltechnologie. Er zijn duizenden talen en dialecten wereldwijd, elk met zijn eigen regels en structuren. Het ontwikkelen van technologie die in al die talen goed functioneert, is complex en vereist veel gespecialiseerde kennis.
- Menselijke taal is uiterst complex, met nuances, ambiguïteit en culturele verschillen die het moeilijk maken voor computers om taal op dezelfde manier te begrijpen als mensen. Die complexiteit maakt taaltechnologie een uitdagend gebied. - Hoewel er veel teksten bestaan die kunnen worden gebruikt om taalmodellen te trainen, vormt dat op zich geen grote uitdaging. Integendeel, het is net die beschikbaarheid van zoveel trainingsdata die taalmodellen als ChatGPT zo goed maakt. - Moderne computers hebben doorgaans voldoende opslagcapaciteit om grote hoeveelheden taaldata te verwerken en op te slaan. Dat is dus geen relevante uitdaging voor taaltechnologie.

2 Hoe het begon
2. Lees de Franse zinnen en hun Nederlandse vertaling.
• Stel je nu voor dat een computerprogramma de Fransen zinnen naar het Nederlands probeert te vertalen door elk Frans woord door zijn Nederlandse vertaling te vervangen.
• Bij welke zinnen gaat zo’n programma de mist in, en waarom?

De volgende elementen vormen een uitdaging voor een vertaalprogramma dat simpelweg woorden door hun vertaling vervangt.

- Woordvolgorde: ‘Comment ça va?’ wordt vertaald door ‘Hoe gaat het?’ en niet door ‘Hoe het gaat?’. Een vertaalprogramma moet dus alle contexten (zoals vragen) in kaart brengen waar de woordvolgorde tussen twee talen verschilt.
- Figuurlijk taalgebruik: cordes betekent letterlijk touwen, maar wordt ook gebruikt in de zegswijze ‘Il pleut des cordes’. Tenzij de twee talen een woord op dezelfde manier figuurlijk gebruiken, moet een vertaalprogramma weten wanneer een woord letterlijk en wanneer het figuurlijk wordt gebruikt. Zelfs als het een lange lijst van uitdrukkingen heeft, bestaan er ongetwijfeld nog voorbeelden van figuurlijk gebruik die niet in die lijst zijn opgenomen.
- Meerdere betekenissen: veel woorden hebben meerdere (letterlijke) betekenissen. Zo kan rapport zowel verstandhouding als verslag betekenen. Een vertaalprogramma moet weten wanneer welke betekenis van toepassing is.
- Sommige woorden worden niet vertaald, zoals des in ‘Il a acheté des cordes.’
Vraag leerlingen die snel de bovenstaande oplossingen vinden ook enkele zinnen te bedenken die zo’n vertaalprogramma niet correct kan vertalen. Dat kan zowel voor het Frans als voor het Engels.

3 De regels van taal

Een van de eerste chatbots, ELIZA, werd gebouwd door de Amerikaan Joseph Weizenbaum in het midden van de jaren 1960. Het programma is vernoemd naar het karakter Eliza Doolittle uit de musical My Fair Lady (1956), die ook

verfilmd werd (in 1964). Daarin leert het personage eveneens hoe ze een gesprek moet voeren. Het programma was eigenlijk als parodie om de oppervlakkigheid van de menselijke communicatie aan te tonen. Weizenbaum was verrast door het bewustzijn dat de eerste gebruikers aan de chatbot toeschreven. Een bekende anekdote daarbij is dat bij een eerste test van het programma zijn secretaresse aan Weizenbaum vroeg om de kamer te verlaten omdat het gesprek te persoonlijk werd. Het eliza-effect, waarbij we onbewust menselijke eigenschappen en drijfveren toeschrijven aan computers, is dan ook naar de ‘geautomatiseerde psychotherapeut’ uit de jaren 1960 genoemd.

4 Sentimentanalyse

3. Beantwoord voor elke tweet de volgende vragen.

● Analyseert Apples software voor sentimentanalyse de tweet als heel negatief, negatief, neutraal, positief of heel positief?
● Is die analyse correct?
● Als de analyse niet correct is, waarom gaat de software in de fout?

1. De nieuwe iPhone is echt geweldig! Analyse: geweldig = positief. Dat is correct.
2. Ongelooflijk hoe vreselijk snel die nieuwe iPhone is! Analyse: vreselijk = negatief. Dat is incorrect. Vreselijk wordt hier namelijk niet als adjectief gebruikt om de iPhone te beschrijven, maar als versterkend bijwoord bij snel, een positieve eigenschap voor een telefoon.
3. De nieuwe iPhone: fantastisch scherm, uitstekende batterijduur. Analyse: fantastisch + uitstekend = heel positief Dat is correct.
4. Nu is het scherm van mijn iPhone alweer kapot. Geweldig! Analyse: geweldig = positief Dat is incorrect. Geweldig wordt hier sarcastisch of ironisch gebruikt. Dat kan het eenvoudige systeem niet detecteren.
5. Ik had verwacht dat de nieuwe iPhone fantastisch zou zijn, maar hij is gewoon goed. Analyse: fantastisch + goed = heel positief Dat is incorrect. De tweet suggereert duidelijk teleurstelling met de nieuwe iPhone, die onze regels niet kunnen vatten.
6. Wil er iemand mijn oude iPhone kopen? Analyse: geen woorden uit het woordenboek = neutraal Dat is correct. © Plantyn
7. Ik heb nog geen slechte ervaring gehad met mijn nieuwe iPhone. Analyse: slechte + geen = positief Dat is correct.
8. De batterijduur van de nieuwe iPhone is echt teleurstellend. Analyse: teleurstellend = negatief Dat is correct.
9. Is het niet vreselijk hoe duur de nieuwste iPhone is? a. Analyse: niet + vreselijk = positief Dat is incorrect. In deze context is niet geen ontkenning van vreselijk. De tweet is negatief.
10. Die nieuwe iPhone! Analyse: geen woorden uit het woordenboek = neutraal Dat is incorrect. Het woordenboek bevat geen emoji’s en kan die dus niet interpreteren.

5 Machines die leren

4. Leer vertalen uit het Zorbisch, een onbestaande taal.

Dit is de vertaling van de Zorbische woorden:
1. Miki = Ik
2. Taso = Jij
3. Bula = eet
4. Dorna = slaap of slaapt
5. Loki = blij zijn
6. Zuma = de hond

Dit is de vertaling van de Zorbische zinnen:

1. Taso dorna = Jij slaapt.
2. Miki bula kafi = Ik eet voer.
3. Loki zuma = De hond is blij.
4. Zuma bula = De hond eet.
5. Taso dorna hapi = Jij slaapt rustig.

Dit zijn de vertalingen van de Nederlandse zinnen naar het Zorbisch:
1. Ik slaap rustig. = Miki dorna hapi
2. Jij eet voer. = Taso bula kafi
3. De hond eet rustig. = Zuma bula hapi

De volgende vertaaloefening laat leerlingen kennismaken met enkele uitdagingen van automatische vertaling, en de grote variatie tussen talen. Daarvoor gebruiken ze een tool als Google Translate https://translate.google.com/) of DeepL (https://www.deepl.com/en/translator). Er is geen account nodig.

Wanneer ze eenvoudige zinnen als ‘Ik ga naar school’ of ‘Ik woon in België’ naar een andere taal en terug laten vertalen, krijgen ze waarschijnlijk dezelfde zin. Dat illustreert hoe goed automatische vertaling werkt. Het slaagt erin de zin correct te vertalen, ook tussen talen die taalkundig heel ver uit elkaar staan.
Voorbeeld (met Google Translate):
- Ik woon in België.
Swahili: Ninaishi Ubelgiji
Nederlands: Ik woon in België.

- Ik ga naar school.
Swahili:Igbo: Anaghị m aga ụlọ akwụkwọ
Nederlands: Ik ga naar school.

Zegswijzen vormen een grotere uitdaging. Dat tonen de volgende voorbeelden aan (met Deepl):

- Nu komt de aap uit de mouw.
Frans: C'est là que le bât blesse.
Nederlands: Daar ligt het probleem.

- Ik heb mijn fiets voor een appel en een ei gekocht.
Frans: J'ai acheté mon vélo pour presque rien.

Nederlands: Ik heb mijn fiets voor bijna niets gekocht.

- Gisteren regende het pijpenstelen.

Frans: Hier, il pleuvait des cordes.
Nederlands: Gisteren regende het pijpenstelen.

- Omdat hij geen Frans sprak, viel hij door de mand.
Frans: Comme il ne parlait pas français, il n'a pas été retenu.
Nederlands: Omdat hij geen Frans sprak, werd hij niet geselecteerd.

In drie van de vier gevallen geeft de vertaaltool een andere Nederlandse zin terug dan de zin waarmee we startten. De verklaring ligt echter niet bij de automatische vertaling. Verschillende talen hebben nu eenmaal andere zegswijzen, zodat een zegswijze vaak vertaald moet worden door zijn letterlijke betekenis. De Nederlandse zinnen die we terugkrijgen betekenen dan ook (min of meer) hetzelfde als de oorspronkelijke zinnen: ze geven gewoon de letterlijke betekenis van de zegswijzen weer. Zo betekent ‘voor een appel en een ei’ ‘bijna niets’. Merk op dat de automatische vertaling de zin ‘gisteren regende het pijpenstelen’ wel behoudt. Dat kan omdat de twee talen in dit geval parallelle zegswijzen hebben (die wel andere woorden gebruiken). In tegenstelling tot de regelgebaseerde systemen uit de sectie hierboven, heeft moderne automatische vertaling geen enkel probleem met de figuurlijke betekenis. Als we experimenteren met talen die verder van het Nederlands staan, stoten we echter wel op de grenzen van automatische vertaling.

Voorbeelden:
- Nu komt de aap uit de mouw.
Japans: □□□□□□□□□□□□□□□
Nederlands: Dit keer kwam de aap uit de tas.

- Ik heb mijn fiets voor een appel en een ei gekocht.
Maori: I hokona e ahau taku paihikara mo tetahi wahi keke.
Nederlands: Ik heb mijn fiets voor een fluitje van een cent verkocht

- Gisteren regende het pijpenstelen.
Tamil: □□□□□□ □□□□□ □□□ □□□□□□□.
Nederlands: Het regende gisteren.

- Omdat hij geen Frans sprak, viel hij door de mand.

Swahili: Kwa sababu hakuzungumza Kifaransa, aliachwa.
Nederlands: Omdat hij geen Frans sprak, bleef hij achter.

De voorbeelden tonen aan dat backtranslation vaak vreemde zinnen oplevert (zoals in zin 1 of 2), of dat de versterkende betekenis van de zegswijze verloren is gegaan (zoals in zin 3). Dat kan aan verschillende factoren te wijten zijn: niet alleen zitten exotische talen vaak taalkundig anders in elkaar dan het Nederlands, meestal zijn er ook minder data beschikbaar om vertaalsoftware op te trainen.

Beide problemen kunnen voor een verlaagde kwaliteit van de vertaling zorgen. Als de leerlingen toch dezelfde zin terugkrijgen (wat bij sommige talenkeuzes kan), laat ze dan een andere taal als tussenstap gebruiken, of zelfs twee talen (bijvoorbeeld Nederlands -> Swahili -> Japans -> Nederlands). Laat snelle leerlingen zelf ook zinnen bedenken met een zegswijze of figuurlijk taalgebruik, en ontdekken welke talen bij backtranslation voor opvallende resultaten zorgen.